Fig. 2. Comparison of models for the distribution of the phone calls duration of a high talkative user, with 3091 calls. TLAC in red, log-normal in green and exponential in black. Visually, for the PDF both the TLAC and the log-normal distribution provide good fits to the CDD but, for the OR, the TLAC clearly provide the best fit.

Given our initial analysis, we may state that the TLAC seems to be a good fit for the CDDs and also serve as an intuitively explanation for how the durations of the calls are generated. However, in order to conclude our answer for Problem 1, we must verify its generality power and also compare it to the log-normal and exponential generality power as well. Thus, we verify which one of the distributions can better fit the CDD of all the users of our dataset that have \( n > 30 \) phone calls. We calculated, for every user, the best fit according to the MLE for the TLAC, the log-normal and exponential distributions and we performed a Kolmogorov-Smirnov goodness of fit test [13], with 5% of significance level, to verify if the user’s CDD is either one of these distributions. For now on, every time we mention that a distribution was correctly fitted, we are implying that we successfully performed a Kolmogorov-Smirnov goodness of fit test.

In Figure 3, we show the percentage of CDDs that could be fitted by a log-normal, a TLAC and a exponential distribution. As we can see, the TLAC distribution can explain the highest fraction of the CDDs and the exponential distribution, the lowest. We observe that the TLAC distribution correctly fit almost 100% of the CDDs for users with \( n < 1000 \). From this point, the quality of the fittings starts to decay, but significantly later than the log-normal distribution. We emphasize that the great majority of users have \( n < 1000 \), what indicates that some of these talkative users’ CDD are probably driven by non natural activities, such as spams, telemarketing or other strong comercial-driven intents. This result, allied to the fact that the TLAC distribution could model more than 96% of the users, make it reasonable to answer Problem 1 claiming that the TLAC distribution is the standard model for CDDs in our dataset.

Finally, we further explore Problem 1 by looking at the OR of the talkative users that were not correctly fitted by the TLAC model. In Figure 4, we show the OR for three of these users and, as we observe, even these customers have a visually good fitting to the TLAC model. These results corroborate even more with the generality power of TLAC. Despite of the fact that the irregularities of these customers’ CDDs unable them to be correctly fitted by the TLAC model, it is clear that the TLAC can represent their CDDs significantly well.
Fig. 3. Percentage of users’ CDDs that were correctly fitted vs. the user’s number of calls $c$. The TLAC distribution is the one that provided better fittings for the whole population of customers with $c > 30$. It correctly fitted more than 96% of the users, only significantly failing to fit users with $c > 10^3$, probably spammers, telemarketers or other non-normal behavior user.

Fig. 4. Odds ratio of 3 talkative customers that were not correctly fitted by the TLAC model

4 TLAC over Time

We know it is trivial to visualize the distribution of users with a determined summarized attribute, such as number of phone calls per month or aggregate calls duration. However, if we want to visualize the distribution and evolution of a temporal feature of the user such as his CDD, things start to get more complicated. Thus, in this section, we tackle the following problem:

Problem 2. EVOLUTION. Given the $\rho_i$ and $\beta_i$ parameters of $N$ customers ($i = 1, 2..., N$), describe how they collectively evolve over time.

We propose two approaches to solve Problem 2. In Section 4.1 we describe the MetaDist solution and, in Section 4.2, we describe the Focal Point approach.

4.1 Group Behavior and Meta-fitting

Since we know that the great majority of users’ CDD can be modeled by the TLAC model, in order to solve Problem 2, we need to figure out how each user $i$ is distributed according to their parameters $\rho_i$ and $\beta_i$ of the TLAC model. If the meta-distribution